DOCUMENT RESUME

ED 463 316 TM 033 748

AUTHOR Osborne, Jason W.; Christianson, William R., II; Gunter, Jason S.

TITLE Educational Psychology from a Statistician's Perspective: A Review of the Quantitative Quality of Our Field.

PUB DATE 2001-04-00

NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Educational Psychology; Educational Research; *Effect Size; Literature Reviews; Reliability; *Research Reports; *Scholarly Journals; *Statistics

IDENTIFIERS *Power (Statistics)

ABSTRACT

The goal of this study was to assess the statistical health of educational psychology literature, both current and past, to: (1) determine the range of effect sizes observed in the current literature (1998-1999); (2) determine the range of observed (or a posteriori) power in the current literature; (3) compare these two statistics to that of the discipline in past years (1939 and 1969); (4) assess the proportion of articles from each of those years reporting testing assumptions of statistical tests, effect size, or power; and (5) assess the reliability of measures used in this research. In all, 55 from 1969 and 96 from the current period were included in these analyses. Results were encouraging, suggesting that most educational psychology research encounters at least moderate (d=0.50) effect sizes, with average power (0.73) that is increasing over that observed in 1969. However, with only 36% of educational psychology studies showing acceptable levels of power (0.80 according to Cohen), only 17% reporting effect sizes, only 8% reporting testing assumptions of statistical tests, and only 2% reporting power, there is still a great deal of room for improvement in the field. (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 463 316

Educational Psychology from a Statistician's Perspective: A Review of the Quantitative Quality of Our Field

Jason W. Osborne
William R. Christianson II
Jason S. Gunter

American Educational Research Association, 2001

TM033748

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

J. Osborne

1

BEST COPY AVAILABLE

Educational Psychology from a Statistician's Perspective: A Review of the Quantitative Quality of Our Field

Jason W. Osborne, William R. Christianson II, and Jason S. Gunter

University of Oklahoma

The goal of this study is to assess the "statistical health" of Educational Psychology literature, both current and past, to determine: (a) the range of effect sizes observed in the current literature (1998-1999), (b) the range of observed (or a posteriori) power in the current literature, (c) compare these two statistics to that of our discipline in past years (1969 and 1939), (d) assess the proportion of articles from each of those years reporting testing assumptions of statistical tests, effect size, and/or power, and (e) assess the reliability of measures used in this research. Results were encouraging, indicating that most EP research encounters at least moderate (d=.50) effect sizes, with average power (.73) that is increasing over that observed at 1969. However, with only 36% of EP studies showing "acceptable" levels of power (.80 according to Cohen), , only 17% reporting effect sizes, only 8% reporting testing assumptions of statistical tests, and only 2% reporting power, there is still a great deal of room for improvement in our field.

Statistical power is the probability of rejecting a null hypothesis. Jacob Cohen (e.g., Cohen, 1962, 1988, 1992) has spent several decades evangelizing the use of power analysis in planning research, reporting research, and particularly in interpreting results where null hypotheses are not rejected. In fact, authors were discussing the issue of power more than half a century ago (e.g., Deemer, 1947). The concept of power is complementary to significance testing and effect size, and, Cohen and others argue, a necessary piece of information in interpreting research results.

Statistical tests tell us the probability of obtaining the observed results given a true null hypothesis. Thus, the ubiquitous $p < .05$ criterion reveals fairly low probability of obtaining observed results by chance or sampling errors, for example. This is an important piece of information to researchers (although those with non-Fisherian leanings would argue that point). However, significance testing is far from sufficient for good research. Most methodologists would argue that effect size reporting is also critical. Effect sizes describe the magnitude of the effect being tested, such as the distance between two means, or the proportion of variance accounted for. As Cohen (1988) and others have pointed out, almost all null hypotheses are ultimately false in an absolute sense (i.e., rarely is a correlation coefficient exactly 0, rarely are two means exactly equal to all decimal places, and so on), and thus, given sufficient power, even the most miniscule effect can produce a $p < .05$. Thus, the reporting of effect sizes, which tells us generally how important an effect is, is crucial.

Authors also argue that even these two pieces of evidence are not sufficient. Power is critical in two different aspects of research. First, Cohen and others argue that no prudent researcher would conduct research without first making a priori analyses to determine the probability of correctly rejecting the null hypothesis. Researchers who fail to do this risk committing Type II errors, or failing to reject a null hypothesis when in fact the null hypothesis should be rejected. This generally occurs when sample sizes are not large enough to reliably detect the expected effect size. Thus, researchers who fail to do power analyses prior to conducting research risk wasting the time and effort spent in conducting research if they do not have sufficient power. Additionally, researchers who fail to do a priori power analyses risk gathering too much data to test their hypotheses—if a power analysis indicates that a sample of one-hundred subjects would be sufficient to detect a particular effect, gathering substantially more is a waste of resources.

Second, a posteriori analyses of power are useful in order to shed light on null results. For example, in a study that fails to reject the null hypothesis, but had power of .90, one can be fairly confident that failing to reject the null was the correct decision. However, when a null hypothesis is not rejected, but there is low power, it is unclear as to whether a Type II error has occurred.

Further, low power has implications for Type I error rates in bodies of literature. As Rossi (1990) argued, low power in a body of literature indicates not only a proliferation of Type II errors in the field, but also the likelihood of a proliferation of Type I errors as well. For example, if the power to detect effects were .20, the probability of correctly rejecting a null hypothesis (.20) is only slightly better than the probability of falsely rejecting a true null (.05). Given the bias in most literatures for publishing significant

results and rejecting null findings, in this case the literature would contain 25% Type I errors. Low power also tends to create other problems in literatures, such as conflicting results in a line of research. For example, in many established bodies of literature there are often studies that find particular effects, and others who fail to find the same effects. If this field has low power, these "conflicting" results could merely be low power in one case leading to a Type II error.

Thus we are left with a situation where statistical power is a very important concept, but reviews of power in many disciplines are discouraging. Cohen's (1962) initial survey of the Journal of Applied and Social Psychology, a top-tier journal at the time, found that power to detect a small effect in this literature was .18, medium effect was .48, and a large effect was .83. In other words, unless researchers in psychology were studying phenomena with large effect sizes, researchers generally had less than a 50-50 chance of detecting effects that existed. Reviews of other areas (see Rossi, 1990; Sedlmeier & Gigerenzer, 1989) paint a bleak picture for more recent research. These reviews indicate that, by the end of the 1980s, little had changed from the early 1960s regarding power.

This study was initiated for several reasons. First, no research has focused on Educational Psychology, and little has been done anywhere in the social sciences since the 1970s or early 1980s. As Cohen (1992) argues, adoption of new methodologies and procedures take a great deal of time (citing the 40 years it took Student's t test to be included in statistical texts). It is possible that in the almost two decades between the more recent reviews of power and the present that the concepts of power have diffused through the field and that the situation has improved. As different sub-disciplines have been shown to vary wildly in power analyses, it is not appropriate to draw conclusions about Educational Psychology research from other psychological disciplines.

Second, previous power surveys have used Cohen's (1962) methodology, where power to detect certain fixed effects are calculated (e.g., effect sizes (d) of 0.2, 0.5, and 0.8). While a decent methodology for equating fields and doing general surveys of power, it is unclear that these accurately reflect effect sizes observed in the field. Nobody knows average effect sizes reported in various fields, and thus, there is no real way to interpret the information provided in previous studies of power. For example, in looking at Cohen's (1962) study, it is unclear as to be horrified (if the average effect size in the field was small ($d = 0.2$) or relieved (if the average effect size in the field was large ($d = 0.8$).

Thus, the prevalent approach to understanding the power of a field, while an important pursuit, is lacking important information to allow for meaningful conclusions. Of course, given the broad diversity of topics studied, methodologies, and analyses even within disciplines like Educational Psychology or Social Psychology, the concept of "power of a field" is even questionable. We argue that, although admittedly imperfect, it is important to, on occasion, examine the field in which we work in order to assess the broad health of the field from a methodological point of view. Just as an individual might make a wholistic judgment about her or his health, regardless of the fact that systems within a single body can vary dramatically in their performance, we as educational psychologists must take a broad measure of the health of the field. This is our general goal for this paper.

More specifically, our goals are to assess the "statistical health" of the field of Educational Psychology by: (a) determining the range of effect sizes observed in the current literature (1998-1999), (b) determine the range of observed (or a posteriori) power in the current literature, (c) compare these two statistics to that of our discipline in past years (1969 and 1939), (d) assess the proportion of articles from each of those years reporting testing assumptions of statistical tests, effect size, and/or power, and (e) assess the reliability of measures used in this research. As many texts and authors argue, not only is power important, but to have healthy research on which be base our inferences, we must have statistics based on tests where assumptions are not grossly violated, researchers should report effect size and power, and should use reliable measures. A literature that fails to incorporate these ingredients is, arguably, of questionable value, as we have insufficient information on which we can draw conclusions.

METHOD

Sample

The unit of analysis for this study, as in previous research on power, is the article. This is to control for the amount of influence any individual article can have over the results. Thus, multiple effects within a study were aggregated to the study level. Further, multi-study articles were similarly aggregated to the article level.

In order to assess the current state of Educational Psychology in a representative fashion, we selected four journals that represent, in our opinion, the best research in the field: Journal of Educational Psychology, Contemporary Educational Psychology, British Journal of Educational Psychology, and

Educational Psychologist. All empirical articles published in these journals during the 1998 and 1999 volumes were surveyed.

To gain a historical context in which to judge the current literature, we decided to survey the volume three decades prior: 1969. This was seven years after the publication of Cohen's seminal article on power, and would allow a comparison of the long-term effects of this work against the short-term effects. Finally, in order to gain a comparison group from before the time power was discussed, we chose the volume three decades prior to that, 1939. As the Journal of Educational Psychology was the only of the four journals that published in those years, and at that time was the primary outlet for high-quality research in the field, that was the only journal surveyed at that time.

Articles containing no statistical analysis (e.g., qualitative, theory, or review articles) or containing statistical analyses for which there are no good methods of computing power (e.g., nonparametric tests, meta-analyses, factor analysis) were excluded from this study. All statistical tests within an article that related to central hypotheses being tested were recorded. Ancillary analyses, such as manipulation checks, and psychometric analyses of measures used, were not recorded as our primary interest is power relating to research hypotheses.

Using methods described in Cohen (1988), all statistics were converted to effect sizes (d), and observed power was computed (as few authors report effect sizes and power, see below). For tests that are not usually associated with d for an effect size (correlation, multiple regression, chi-square, e.g.), algebraic manipulations derived from Cohen (1988) were called upon to transform all other effect size indices to ds for comparability. For all tests, alpha of .05 was assumed, and where appropriate, two-tailed tests were assumed as well. The number of relevant effect sizes in individual articles ranged from one to fifty-four. All effect size and power information were aggregated to the article level. Finally, to allow for comparison with other power surveys, power to detect small (d=.2), medium (d=.5), and large (d=.8) effects were calculated.

Other information gathered from articles include whether the research was an experimental or correlational design, used a college sample or not, and whether the authors reported effect sizes, observed or a priori power, and whether the authors gave any indication that the assumptions of the analyses used were checked. Finally, reliability information on measures used was gathered, but have not yet been analyzed.

RESULTS AND DISCUSSION

Due to time constraints, all articles were not able to be included in this presentation. In total, 19 from 1939, 55 from 1969, and 96 from the present (1998-1999) were included in the database. Due to the small sample from 1939, those articles were not included in the analyses.

Effect size

In examining articles from the present (1998-1999) there was a significant differences in observed effect size across journals (average d for the British Journal of Educational Psychology (d=.57) was significantly lower than the others, who ranged from d=.73 to d=.80, $p < .03$).

The average effect size for Educational Psychology articles in the present was d=.73 (SD=.32), while the average effect size from 1969 was d=.69 (SD=.33). There was no significant difference between the two groups. It is interesting to note that the average effect size for this literature has been consistently near what Cohen termed "large" for the last 30 years. Although there is great variation (a 95% confidence interval for current articles leaves a range of d=.10 to d=1.36), this is impressive as most authors assume social science research produces and deals with relatively small effect sizes. 76% of modern Educational Psychology research deals with effect sizes of at least medium size (d=.50).

Power

There were significant differences in observed power across the journal articles for studies from the present (average power for BJEP was .65, for JEP was .76, the rest in-between, $p < .01$). While there were no significant differences in effect size across years, observed power in studies has improved significantly from 1969 (power = .63, SD=.18) to the present (power=.73, SD=.21, $p < .002$).

When examining the power to detect small (d=.2), medium (d=.5) or large (d=.8) effects showed similar increases:

Power to detect:	1969	1999	p <
small	.20	.27	.04
medium	.60	.71	.03
large	.84	.89	.08

Most of this can be attributed to larger sample sizes, facilitated by easy access to computing that was lacking in 1969. Average cell sizes for 1969 was N= 82.6, while for the present was N=152.5.

While computing average power is fine, averages can be substantially influenced by outliers. Of more interest was the number of studies that showed acceptable levels of power. According to Cohen and others, the minimum acceptable level of power researchers should accept is .80, indicating an 80% chance of rejecting a true null hypothesis at a given level of effect. Thus, the percentage of studies from each year meeting the power = .80 criterion was calculated.

	1969	present
Non-experimental	4%	44%
experimental	21%	29%

There was a significant main effect for year, in that only 13% of studies from 1969 had what Cohen identified as sufficient power. In the present studies 36% of the studies had sufficient power ($p<.01$). This effect was qualified by an interaction with whether the study was an experimental or non-experimental study (an experimental study was operationally defined here as a study with random assignment to condition and active manipulation of an independent variable). It is clear that it is the non-experimental research that has had tremendous gains in power over the last three decades. While these proportions are troublingly low, they are increasing.

Thus, while authors (e.g., Rossi, 1990; Sedlmeier & Gigerenzer, 1989) have argued that power in general has not changed significantly since Cohen's first power analysis in 1962, at least for Educational Psychology it does appear that power is improving.

Other indicators of high-quality research

During our survey of the literature, we gathered data on several other variables that are related to the production of a high-quality literature, such as reliability of measures used, whether assumptions of statistical tests were tested, and whether effect sizes and power were reported in the article. Additionally, we tracked whether studies used experimental methodology or not. While not directly related to the quality of a literature, some mix of experimental and non-experimental methodology is probably desirable, as experimental methodology often allows for stronger inference than non-experimental methodology. Finally, as the social sciences have been tremendously reliant on college (often introduction to psychology) subject pools, we also tracked whether the study was done on college samples or not. Again, while not necessarily a negative thing, there is probably some virtue to having a mix of college and non-college samples in the literature, as much of Educational Psychology is aimed at K-12 populations.

	1969	Present	p <
Report reliability	15%	26%	.10
True Experiment	60%	15%	.0001
College sample	36%	23%	.07
Reported effect size	3%	17%	.02
Reported power	0%	2%	*ns*
Reported testing assumptions	7%	8%	*ns*

Several interesting trends emerge from these data. First, we were surprised to see few studies reporting the reliability of their measures. While it is possible that many were behavioral observations or using measures of undetermined reliability, these statistics suggest that, as a field, we are not paying attention to measurement to the extent we should, particularly since without reliable and valid measurement we have no data to discuss. For those reporting reliability, it was generally in the acceptable range (average for 1969 was .86, for the present was .83).

Second, not surprisingly, the Educational Psychology of the 1960s was largely experimental. Present-day research is primarily non-experimental, with less than one in seven studies using experimental paradigms. This may be undesirable for the field as a whole. Although there are problems with experimental paradigms, there are strengths as well, and we would probably not want to see Educational Psychology become exclusively non-experimental.

Third, there has been a move away from relying on college-age samples. In our opinion, this is a positive move, as college undergraduates are rarely representative of the populations to which we desire to generalize. Indeed, the undergraduates in these studies are probably not even representative of college undergraduates in general, as they tend to be underclassmen in psychology or related social science disciplines.

Fourth, while there is a significant increase in the proportion of studies reporting effect sizes, this proportion is too low. Many authors have argued that significance level is not sufficient information for interpretation of effects, and the APA has moved to require effect size reporting, largely in response to the perceived desirability of reporting these statistics and the lack of authors actually reporting them.

Fifth, almost no studies report observed or *a priori* power for statistics tests. This is tremendously problematic, as knowing the power of tests has many benefits to the reader (e.g., interpretation of non-significant results) as well as the researcher (reducing Type II errors). This is something authors and editors need to remedy in order to increase the quality of the literature. Although average power in the field is higher than most other reviews indicate, only one-third of all studies currently published in top-tier journals have sufficient power to test their hypotheses. This raises some serious issues with the quality of the conclusions we can draw on the basis of the existing literature.

Finally, only 7% and 8% of authors report having tested assumptions of the statistical tests they utilize. This may be due to mixed signals researchers get from statistics texts and statisticians. On one hand, many texts report tests as being generally robust to moderate violations of assumptions. On the other hand, even simple issues such as having outliers or fringeliers in a data set can dramatically alter statistics and the probability of a Type I or Type II error (Osborne & Holt, 2001). Thus, while many tests are robust to certain types of violations, that does not mean that researchers should not test or report having tested the assumptions relating to their analytic procedures.

SUMMARY

There is both good and bad news for Educational Psychology as a field. We posit that there are several requirements to a high-quality scientific literature: sufficient power to detect effects and avoid high proportions of Type I and Type II errors in the literature, reporting of effect size and power statistics to enhance interpretation of results, reporting testing of assumptions of analytic methods, reduced reliance on college underclass students for research, and a mix of experimental and non-experimental research.

Observed effect sizes in Educational Psychology are larger than we expected. While there is great variation, most studies in this literature report at least moderate effect sizes (d= .5). Further, average power in the current literature is .73. While this is not ideal, and leaves room for many Type II errors, it does help to keep Type I error rates at relatively low levels (probably about 6.8%). Indications of increasing power from 1969-1999 is also positive. However, with only 36% of current studies showing observed power in excess of .80, there is still a great deal of room for improvement.

Finally, the other indicators of quality research are similarly mixed. For example, with almost 75% of current authors failing to report reliability, only 17% reporting effect sizes, 2% reporting power, and 8% reporting testing assumptions of their analyses, it would be fair to raise some serious questions as to exactly what sorts of conclusions we feel comfortable drawing from this literature. On the other hand, a decreasing reliance on college samples bodes well for the external validity of our research as a whole.

REFERENCES:

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (second edition). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Deemer, W. L. (1947). The power of the t test and the estimation of required sample size. Journal of Educational Psychology, 38, 329-342.

Osborne, J. W., & Holt, A. (2001). The effect of outliers and fringeliers on the accuracy and error rate of quantitative analyses. Unpublished paper, University of Oklahoma.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? Journal of Counseling and Clinical Psychology, 58, 646-656.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

AUTHOR NOTES

We would like to thank Cheryl Murdock for help with data extraction. The first author can be contacted at the Department of Educational Psychology, University of Oklahoma, 820 Van Vleet Oval, Norman, OK, 73019 or via email at josborne@ou.edu.

Educational Psychology from a Statistician's Perspective: A Review of the Quantitative Quality of Our Field

Jason W. Osborne, William R. Christianson II, and Jason S. Gunter

University of Oklahoma

The goal of this study is to assess the "statistical health" of Educational Psychology literature, both current and past, to determine: (a) the range of effect sizes observed in the current literature (1998-1999), (b) the range of observed (or a posteriori) power in the current literature, (c) compare these two statistics to that of our discipline in past years (1969 and 1939), (d) assess the proportion of articles from each of those years reporting testing assumptions of statistical tests, effect size, and/or power, and (e) assess the reliability of measures used in this research. Results were encouraging, indicating that most EP research encounters at least moderate (d=.50) effect sizes, with average power (.73) that is increasing over that observed at 1969. However, with only 36% of EP studies showing "acceptable" levels of power (.80 according to Cohen), , only 17% reporting effect sizes, only 8% reporting testing assumptions of statistical tests, and only 2% reporting power, there is still a great deal of room for improvement in our field.

Statistical power is the probability of rejecting a null hypothesis. Jacob Cohen (e.g., Cohen, 1962, 1988, 1992) has spent several decades evangelizing the use of power analysis in planning research, reporting research, and particularly in interpreting results where null hypotheses are not rejected. In fact, authors were discussing the issue of power more than half a century ago (e.g., Deemer, 1947). The concept of power is complementary to significance testing and effect size, and, Cohen and others argue, a necessary piece of information in interpreting research results.

Statistical tests tell us the probability of obtaining the observed results given a true null hypothesis. Thus, the ubiquitous $p < .05$ criterion reveals fairly low probability of obtaining observed results by chance or sampling errors, for example. This is an important piece of information to researchers (although those with non-Fisherian leanings would argue that point). However, significance testing is far from sufficient for good research. Most methodologists would argue that effect size reporting is also critical. Effect sizes describe the magnitude of the effect being tested, such as the distance between two means, or the proportion of variance accounted for. As Cohen (1988) and others have pointed out, almost all null hypotheses are ultimately false in an absolute sense (i.e., rarely is a correlation coefficient exactly 0, rarely are two means exactly equal to all decimal places, and so on), and thus, given sufficient power, even the most miniscule effect can produce a $p < .05$. Thus, the reporting of effect sizes, which tells us generally how important an effect is, is crucial.

Authors also argue that even these two pieces of evidence are not sufficient. Power is critical in two different aspects of research. First, Cohen and others argue that no prudent researcher would conduct research without first making a priori analyses to determine the probability of correctly rejecting the null hypothesis. Researchers who fail to do this risk committing Type II errors, or failing to reject a null hypothesis when in fact the null hypothesis should be rejected. This generally occurs when sample sizes are not large enough to reliably detect the expected effect size. Thus, researchers who fail to do power analyses prior to conducting research risk wasting the time and effort spent in conducting research if they do not have sufficient power. Additionally, researchers who fail to do a priori power analyses risk gathering too much data to test their hypotheses—if a power analysis indicates that a sample of one-hundred subjects would be sufficient to detect a particular effect, gathering substantially more is a waste of resources.

Second, a posteriori analyses of power are useful in order to shed light on null results. For example, in a study that fails to reject the null hypothesis, but had power of .90, one can be fairly confident that failing to reject the null was the correct decision. However, when a null hypothesis is not rejected, but there is low power, it is unclear as to whether a Type II error has occurred.

Further, low power has implications for Type I error rates in bodies of literature. As Rossi (1990) argued, low power in a body of literature indicates not only a proliferation of Type II errors in the field, but also the likelihood of a proliferation of Type I errors as well. For example, if the power to detect effects were .20, the probability of correctly rejecting a null hypothesis (.20) is only slightly better than the probability of falsely rejecting a true null (.05). Given the bias in most literatures for publishing significant

results and rejecting null findings, in this case the literature would contain 25% Type I errors. Low power also tends to create other problems in literatures, such as conflicting results in a line of research. For example, in many established bodies of literature there are often studies that find particular effects, and others who fail to find the same effects. If this field has low power, these "conflicting" results could merely be low power in one case leading to a Type II error.

Thus we are left with a situation where statistical power is a very important concept, but reviews of power in many disciplines are discouraging. Cohen's (1962) initial survey of the Journal of Applied and Social Psychology, a top-tier journal at the time, found that power to detect a small effect in this literature was .18, medium effect was .48, and a large effect was .83. In other words, unless researchers in psychology were studying phenomena with large effect sizes, researchers generally had less than a 50-50 chance of detecting effects that existed. Reviews of other areas (see Rossi, 1990; Sedlmeier & Gigerenzer, 1989) paint a bleak picture for more recent research. These reviews indicate that, by the end of the 1980s, little had changed from the early 1960s regarding power.

This study was initiated for several reasons. First, no research has focused on Educational Psychology, and little has been done anywhere in the social sciences since the 1970s or early 1980s. As Cohen (1992) argues, adoption of new methodologies and procedures take a great deal of time (citing the 40 years it took Student's t test to be included in statistical texts). It is possible that in the almost two decades between the more recent reviews of power and the present that the concepts of power have diffused through the field and that the situation has improved. As different sub-disciplines have been shown to vary wildly in power analyses, it is not appropriate to draw conclusions about Educational Psychology research from other psychological disciplines.

Second, previous power surveys have used Cohen's (1962) methodology, where power to detect certain fixed effects are calculated (e.g., effect sizes (d) of 0.2, 0.5, and 0.8). While a decent methodology for equating fields and doing general surveys of power, it is unclear that these accurately reflect effect sizes observed in the field. Nobody knows average effect sizes reported in various fields, and thus, there is no real way to interpret the information provided in previous studies of power. For example, in looking at Cohen's (1962) study, it is unclear as to be horrified (if the average effect size in the field was small (d = 0.2) or relieved (if the average effect size in the field was large (d = 0.8).

Thus, the prevalent approach to understanding the power of a field, while an important pursuit, is lacking important information to allow for meaningful conclusions. Of course, given the broad diversity of topics studied, methodologies, and analyses even within disciplines like Educational Psychology or Social Psychology, the concept of "power of a field" is even questionable. We argue that, although admittedly imperfect, it is important to, on occasion, examine the field in which we work in order to assess the broad health of the field from a methodological point of view. Just as an individual might make a wholistic judgment about her or his health, regardless of the fact that systems within a single body can vary dramatically in their performance, we as educational psychologists must take a broad measure of the health of the field. This is our general goal for this paper.

More specifically, our goals are to assess the "statistical health" of the field of Educational Psychology by: (a) determining the range of effect sizes observed in the current literature (1998-1999), (b) determine the range of observed (or a posteriori) power in the current literature, (c) compare these two statistics to that of our discipline in past years (1969 and 1939), (d) assess the proportion of articles from each of those years reporting testing assumptions of statistical tests, effect size, and/or power, and (e) assess the reliability of measures used in this research. As many texts and authors argue, not only is power important, but to have healthy research on which be base our inferences, we must have statistics based on tests where assumptions are not grossly violated, researchers should report effect size and power, and should use reliable measures. A literature that fails to incorporate these ingredients is, arguably, of questionable value, as we have insufficient information on which we can draw conclusions.

METHOD

Sample

The unit of analysis for this study, as in previous research on power, is the article. This is to control for the amount of influence any individual article can have over the results. Thus, multiple effects within a study were aggregated to the study level. Further, multi-study articles were similarly aggregated to the article level.

In order to assess the current state of Educational Psychology in a representative fashion, we selected four journals that represent, in our opinion, the best research in the field: Journal of Educational Psychology, Contemporary Educational Psychology, British Journal of Educational Psychology, and

Educational Psychologist. All empirical articles published in these journals during the 1998 and 1999 volumes were surveyed.

To gain a historical context in which to judge the current literature, we decided to survey the volume three decades prior: 1969. This was seven years after the publication of Cohen's seminal article on power, and would allow a comparison of the long-term effects of this work against the short-term effects. Finally, in order to gain a comparison group from before the time power was discussed, we chose the volume three decades prior to that, 1939. As the Journal of Educational Psychology was the only of the four journals that published in those years, and at that time was the primary outlet for high-quality research in the field, that was the only journal surveyed at that time.

Articles containing no statistical analysis (e.g., qualitative, theory, or review articles) or containing statistical analyses for which there are no good methods of computing power (e.g., nonparametric tests, meta-analyses, factor analysis) were excluded from this study. All statistical tests within an article that related to central hypotheses being tested were recorded. Ancillary analyses, such as manipulation checks, and psychometric analyses of measures used, were not recorded as our primary interest is power relating to research hypotheses.

Using methods described in Cohen (1988), all statistics were converted to effect sizes (d), and observed power was computed (as few authors report effect sizes and power, see below). For tests that are not usually associated with d for an effect size (correlation, multiple regression, chi-square, e.g.), algebraic manipulations derived from Cohen (1988) were called upon to transform all other effect size indices to ds for comparability. For all tests, alpha of .05 was assumed, and where appropriate, two-tailed tests were assumed as well. The number of relevant effect sizes in individual articles ranged from one to fifty-four. All effect size and power information were aggregated to the article level. Finally, to allow for comparison with other power surveys, power to detect small (d=.2), medium (d=.5), and large (d=.8) effects were calculated.

Other information gathered from articles include whether the research was an experimental or correlational design, used a college sample or not, and whether the authors reported effect sizes, observed or a priori power, and whether the authors gave any indication that the assumptions of the analyses used were checked. Finally, reliability information on measures used was gathered, but have not yet been analyzed.

RESULTS AND DISCUSSION

Due to time constraints, all articles were not able to be included in this presentation. In total, 19 from 1939, 55 from 1969, and 96 from the present (1998-1999) were included in the database. Due to the small sample from 1939, those articles were not included in the analyses.

Effect size

In examining articles from the present (1998-1999) there was a significant differences in observed effect size across journals (average d for the British Journal of Educational Psychology (d=.57) was significantly lower than the others, who ranged from d=.73 to d=.80, $p < .03$).

The average effect size for Educational Psychology articles in the present was d=.73 (SD=.32), while the average effect size from 1969 was d=.69 (SD=.33). There was no significant difference between the two groups. It is interesting to note that the average effect size for this literature has been consistently near what Cohen termed "large" for the last 30 years. Although there is great variation (a 95% confidence interval for current articles leaves a range of d=.10 to d=1.36), this is impressive as most authors assume social science research produces and deals with relatively small effect sizes. 76% of modern Educational Psychology research deals with effect sizes of at least medium size (d=.50).

Power

There were significant differences in observed power across the journal articles for studies from the present (average power for BJEP was .65, for JEP was .76, the rest in-between, $p < .01$). While there were no significant differences in effect size across years, observed power in studies has improved significantly from 1969 (power = .63, SD=.18) to the present (power=.73, SD=.21, $p < .002$).

When examining the power to detect small (d=.2), medium (d=.5) or large (d=.8) effects showed similar increases:

Power to detect:	1969	1999	p <
small	.20	.27	.04
medium	.60	.71	.03
large	.84	.89	.08

Most of this can be attributed to larger sample sizes, facilitated by easy access to computing that was lacking in 1969. Average cell sizes for 1969 was N= 82.6, while for the present was N=152.5.

While computing average power is fine, averages can be substantially influenced by outliers. Of more interest was the number of studies that showed acceptable levels of power. According to Cohen and others, the minimum acceptable level of power researchers should accept is .80, indicating an 80% chance of rejecting a true null hypothesis at a given level of effect. Thus, the percentage of studies from each year meeting the power = .80 criterion was calculated.

	1969	present
Non-experimental	4%	44%
experimental	21%	29%

There was a significant main effect for year, in that only 13% of studies from 1969 had what Cohen identified as sufficient power. In the present studies 36% of the studies had sufficient power ($p<.01$). This effect was qualified by an interaction with whether the study was an experimental or non-experimental study (an experimental study was operationally defined here as a study with random assignment to condition and active manipulation of an independent variable). It is clear that it is the non-experimental research that has had tremendous gains in power over the last three decades. While these proportions are troublingly low, they are increasing.

Thus, while authors (e.g., Rossi, 1990; Sedlmeier & Gigerenzer, 1989) have argued that power in general has not changed significantly since Cohen's first power analysis in 1962, at least for Educational Psychology it does appear that power is improving.

Other indicators of high-quality research

During our survey of the literature, we gathered data on several other variables that are related to the production of a high-quality literature, such as reliability of measures used, whether assumptions of statistical tests were tested, and whether effect sizes and power were reported in the article. Additionally, we tracked whether studies used experimental methodology or not. While not directly related to the quality of a literature, some mix of experimental and non-experimental methodology is probably desirable, as experimental methodology often allows for stronger inference than non-experimental methodology. Finally, as the social sciences have been tremendously reliant on college (often introduction to psychology) subject pools, we also tracked whether the study was done on college samples or not. Again, while not necessarily a negative thing, there is probably some virtue to having a mix of college and non-college samples in the literature, as much of Educational Psychology is aimed at K-12 populations.

	1969	Present	p <
Report reliability	15%	26%	.10
True Experiment	60%	15%	.0001
College sample	36%	23%	.07
Reported effect size	3%	17%	.02
Reported power	0%	2%	*ns*
Reported testing assumptions	7%	8%	*ns*

Several interesting trends emerge from these data. First, we were surprised to see few studies reporting the reliability of their measures. While it is possible that many were behavioral observations or using measures of undetermined reliability, these statistics suggest that, as a field, we are not paying attention to measurement to the extent we should, particularly since without reliable and valid measurement we have no data to discuss. For those reporting reliability, it was generally in the acceptable range (average for 1969 was .86, for the present was .83).

Second, not surprisingly, the Educational Psychology of the 1960s was largely experimental. Present-day research is primarily non-experimental, with less than one in seven studies using experimental paradigms. This may be undesirable for the field as a whole. Although there are problems with experimental paradigms, there are strengths as well, and we would probably not want to see Educational Psychology become exclusively non-experimental.

Third, there has been a move away from relying on college-age samples. In our opinion, this is a positive move, as college undergraduates are rarely representative of the populations to which we desire to generalize. Indeed, the undergraduates in these studies are probably not even representative of college undergraduates in general, as they tend to be underclassmen in psychology or related social science disciplines.

Fourth, while there is a significant increase in the proportion of studies reporting effect sizes, this proportion is too low. Many authors have argued that significance level is not sufficient information for interpretation of effects, and the APA has moved to require effect size reporting, largely in response to the perceived desirability of reporting these statistics and the lack of authors actually reporting them.

Fifth, almost no studies report observed or *a priori* power for statistics tests. This is tremendously problematic, as knowing the power of tests has many benefits to the reader (e.g., interpretation of non-significant results) as well as the researcher (reducing Type II errors). This is something authors and editors need to remedy in order to increase the quality of the literature. Although average power in the field is higher than most other reviews indicate, only one-third of all studies currently published in top-tier journals have sufficient power to test their hypotheses. This raises some serious issues with the quality of the conclusions we can draw on the basis of the existing literature.

Finally, only 7% and 8% of authors report having tested assumptions of the statistical tests they utilize. This may be due to mixed signals researchers get from statistics texts and statisticians. On one hand, many texts report tests as being generally robust to moderate violations of assumptions. On the other hand, even simple issues such as having outliers or fringeliers in a data set can dramatically alter statistics and the probability of a Type I or Type II error (Osborne & Holt, 2001). Thus, while many tests are robust to certain types of violations, that does not mean that researchers should not test or report having tested the assumptions relating to their analytic procedures.

SUMMARY

There is both good and bad news for Educational Psychology as a field. We posit that there are several requirements to a high-quality scientific literature: sufficient power to detect effects and avoid high proportions of Type I and Type II errors in the literature, reporting of effect size and power statistics to enhance interpretation of results, reporting testing of assumptions of analytic methods, reduced reliance on college underclass students for research, and a mix of experimental and non-experimental research.

Observed effect sizes in Educational Psychology are larger than we expected. While there is great variation, most studies in this literature report at least moderate effect sizes (d= .5). Further, average power in the current literature is .73. While this is not ideal, and leaves room for many Type II errors, it does help to keep Type I error rates at relatively low levels (probably about 6.8%). Indications of increasing power from 1969-1999 is also positive. However, with only 36% of current studies showing observed power in excess of .80, there is still a great deal of room for improvement.

Finally, the other indicators of quality research are similarly mixed. For example, with almost 75% of current authors failing to report reliability, only 17% reporting effect sizes, 2% reporting power, and 8% reporting testing assumptions of their analyses, it would be fair to raise some serious questions as to exactly what sorts of conclusions we feel comfortable drawing from this literature. On the other hand, a decreasing reliance on college samples bodes well for the external validity of our research as a whole.

REFERENCES:

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (second edition). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Deemer, W. L. (1947). The power of the t test and the estimation of required sample size. Journal of Educational Psychology, 38, 329-342.

Osborne, J. W., & Holt, A. (2001). The effect of outliers and fringeliers on the accuracy and error rate of quantitative analyses. Unpublished paper, University of Oklahoma.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? Journal of Counseling and Clinical Psychology, 58, 646-656.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

AUTHOR NOTES

We would like to thank Cheryl Murdock for help with data extraction. The first author can be contacted at the Department of Educational Psychology, University of Oklahoma, 820 Van Vleet Oval, Norman, OK, 73019 or via email at josborne@ou.edu.